8A

УДК 004.272.2

Н.Г. Аксак, А.Ю. Лебёдкина
Харьковский национальный университет радиоэлектроники, г. Харьков, Украина
axak@kture.kharkov.ua

Методы и модели производительности обучения многослойных нейронных сетей в распределенных компьютерных средах

В работе предложены методы и модели производительности параллельных процедур, которые позволяют эффективно применять многослойную нейронную сеть в распределенных компьютерных средах с различными топологиями («решетка», «полносвязный граф», «звезда»). Достоверность использования предложенных методов и моделей подтверждается экспериментальными исследованиями.

Введение

Стремительное развитие высокопроизводительных вычислительных систем повлекло за собой множество следствий. Появилась возможность параллельно решать сложные прикладные задачи с большим объемом вычислений. Для этого необходимы специальные технологии и методы, допускающие возможность разделения задач на фрагменты, которые можно выполнять независимо друг от друга. Примерами решения больших задач в распределенных вычислительных средах являются [1]:

– в Центре «Биоинженерия» РАН определена скрытая периодичность в генетических последовательностях. Задача решена за 63 часа (на 1 CPU требуется 2 года), участвовало 8 городов,10 организаций, 14 кластеров, 407 CPUs;

– в ПензГУ решена задача дифракции электромагнитного поля на диэлектрическом анизотропном теле произвольной формы. Потребовалось 26 дней на фоне работы СКЦ, в то время как на 1 CPU – 4 года. Использовалось 4 кластера СКЦ НИВЦ МГУ;

– в ИБМХ РАМН, Гематологический центр РАМН осуществлен поиск молекул-ингибиторов для заданных белков-мишеней (тромбин). Задействовано было 270 CPUs, решена за 11 дней (на 1 CPU требуется 4,5 года), участвовало 2 города, 3 кластера в учебном классе.

В то же время для решения неформализованных задач широкое распространение получили искусственные нейронные сети, при этом их естественная внутренняя структура позволяет реализовывать параллельное обучение и функционирование нейросетевых процедур. Однако на сегодняшний день основные нейросетевые парадигмы предназначены для моделирования на традиционных однопроцессорных компьютерах. Кроме того, время обучения нейронных сетей при больших объемах исходных данных, обрабатываемых последовательно, может быть очень велико.

Таким образом, актуальными являются синтез и адаптация нейронных сетей на высокопроизводительные вычислительные архитектуры с наименьшим временем обучения.

Целью данной статьи является разработка методов и моделей оценивания производительности распределенных нейропроцедур со сверхбольшим объемом данных с учетом адаптации на высокопроизводительные архитектуры.

Аксак Н.Г., Лебедкина А.Ю.

Методы оценивания эффективности работы параллельной процедуры

Под распределенными вычислениями понимается способ решения трудоёмких вычислительных задач с использованием нескольких компьютеров, объединённых в параллельную вычислительную систему.

Оценивание эффективности распределенной процедуры осуществляется в зависимости от количества доступных вычислителей, соответствующих значению P, которые идентифицируются номерами $r=\overline{0,P-1}$, где $r=0$ – управляющий вычислитель, $r=\overline{1,P-1}$ – рабочие вычислители.

Сетевой закон Амдала [2], [3] является традиционным методом определения теоретического ускорения распределенной процедуры

$$S=\frac{1}{a+\frac{1-a}{P}+c}=\frac{1}{\frac{Q_{пос}}{Q}+\frac{1-Q_{пос}/Q}{P}+\frac{Q_c t_c}{Qt}}=\frac{Q}{Q_{пос}+\frac{Q_{пар}}{p}+Q_c\frac{t_c}{t}}=\frac{Q}{Q_p+Q_c\frac{t_c}{t}}, \quad (1)$$

где P – количество вычислителей,

Q – общее количество тактов выполнения последовательной процедуры,

$Q_{пос}$ – количество скалярных непараллельных операций в параллельной процедуре,

$a=\frac{Q_{пос}}{Q}$ – удельный вес последовательных операций в параллельной процедуре,

$Q_{пар}$ – количество скалярных параллельных операций в параллельной процедуре,

$Q_p=Q_{пос}+\frac{Q_{пар}}{P}$ – общее количество тактов выполнения параллельной процедуры,

$c=\frac{Q_c t_c}{Qt}$ – коэффициент сетевой деградации вычислений,

Q_c – общее количество передач данных,

t_c – пропускная способность сети,

t – пиковая производительность вычислителя.

Однако закон Амдала при определении общего количества тактов выполнения параллельной процедуры $Q_p=Q_{пос}+\frac{Q_{пар}}{P}$ предполагает в качестве обязательного условия равномерное распределение объемов данных каждому вычислителю. Таким образом, графическая интерпретация закона Амдала представляет собой непрерывный прирост производительности до некоторой точки максимума, что расходится с реальной производительностью процедуры с учетом синхронизации параллельных процессов.

Тогда метод оценивания эффективности распределенной процедуры по подчиненному принципу «master/slave» при параллелизме на уровне задач будет иметь вид

$$S=\frac{Q\times t^0}{\max\limits_{r=\overline{0,P-1}} Q^r_{пар}t^r+Q_{пос}\times t^0+V+Q_c\times D\times t_c}, \quad (2)$$

где $t^r=F^r\times N^r_c\times N^r_o$, $r=\overline{0,P-1}$ – пиковая производительность r-о вычислителя,

F^r, $r=\overline{0,P-1}$ – тактовая частота r-го вычислителя,

N^r_c, $r=\overline{0,P-1}$ – количество вычислительных ядер r-го вычислителя,

N^r_o, $r=\overline{0,P-1}$ – количество операций с плавающей запятой на такт r-го вычислителя,

$Q^{r}_{пар}$, $r=\overline{0,P-1}$ – количество скалярных параллельных операций, выполняющихся на r-ом вычислителе,

$Q_{пос}$ – количество скалярных последовательных операций, выполняющихся на управляющем вычислителе,

D – диаметр, определяющий максимальное расстояние между двумя вычислителями сети,

V – латентность сети.

При параллелизме на уровне данных метод оценивания эффективности распределенной процедуры принимает вид

$$S=\frac{Q\times t^{0}}{\max\limits_{r=\overline{0,P-1}} Q^{r}_{t}t^{r}+Q_{пос}\times t^{0}+V+Q_{c}\times D\times t_{c}}, \qquad (3)$$

где Q_{u} – количество скалярных последовательных операций на один параллельный такт,

K_{r}, $r=\overline{0,P-1}$ – критерий равномерного распределения параллельных операций в зависимости от номера вычислителя r, определяемый в соответствии с (4), (5),

$Q^{r}_{t}=K_{r}Q_{u}$, $r=\overline{0,P-1}$ – количество тактов выполнения параллельных операций, выполняющихся на r-ом вычислителе.

Максимально возможные значения критерия равномерного распределения параллельных операций K_{max} определяются как

$$K_{max}=\begin{cases}\left[Q_{пар}/P\right], & Q_{пар}\,\vdots\,P,\\ \left[Q_{пар}/P\right]+1, & Q_{пар}\,\not\vdots\,P.\end{cases} \qquad (4)$$

Текущие значения критерия K_{r} определяются в зависимости от номера вычислителя

$$K_{r}=\begin{cases}\left\{(r-1)\left(\left[Q_{пар}/P\right]+1\right)+\varphi\right\}, \ \varphi=\overline{1,\left(\left[Q_{пар}/P\right]+1\right)}, & 1\le r\le r_{d},\\ \left\{(P-b)\left(\left[Q_{пар}/P\right]+1\right)+(r-P+b-1)\left[Q_{пар}/P\right]+\varphi\right\}, \ \varphi=\overline{1,\left[Q_{пар}/P\right]}, & r>r_{d},\end{cases} \qquad (5)$$

где $r_{d}=P-b$ – номер вычислителя, начиная с которого уменьшается на единицу значение критерия равномерного распределения параллельных операций K_{r},

$b=Q_{пар}\times \mathrm{mod}(P-1)$ – количество вычислителей с большей нагрузкой.

В выражениях (2), (3) значение Q_{p} определяется с учетом барьерной синхронизации параллельных процессов как при гетерогенной, так и при гомогенной распределенной среде, что увеличивает точность оценивания эффективности работы параллельной процедуры.

Модели производительности процедуры распределенного обучения многослойной нейронной сети

Одним из преимуществ нейронных сетей является возможность обучения, которое заключается в нахождении коэффициентов связей между нейронами [4].

Обозначим через $S(U,I,T,L,P,n_1,n_2,\ldots,n_L)$ ускорение задачи распределенного обучения L-слойной нейронной сети $(n_1-n_2-\ldots-n_L)$, где $U-$ количество эпох обучения, I и $T-$ соответственно количество примеров в обучающей и тестовой выборках, P – число доступных вычислителей, которые идентифицируются номерами $r=\overline{0,P-1}$. На основании выражения (3) построим параметрическую модель производительности распределенного обучения многослойной нейронной сети

$$S(U,I,T,L,P,n_1,n_2,\ldots,n_L)=\frac{Q(U,I,T,L,n_1,n_2,\ldots,n_L)\times t^0}{Q_p(U,I,T,L,n_1,n_2,\ldots,n_L,r)+V+Q_c(U,I,T,L,P)\times D\times t_c}, \qquad (6)$$

где алгоритмические составляющие Q_p, Q_c для модели производительности распределенного обучения многослойной нейронной сети с топологией передачи данных «полносвязный граф» соответственно определены как

$$\left.\begin{aligned}
Q_p = U\times(I\times(&\sum_{m=2}^{L}\max_{r=\overline{0,P-1}}(K_{m,r}t^r(2n_{m-1}+3))+\sum_{m=2}^{L-1}\max_{r=\overline{0,P-1}}(K_{m,r}t^r(3+2n_{m+1})n_{m-1})+\\
&+5\max_{r=\overline{0,P-1}}(K_{L,r}t^r n_{L-1}))+T\times(\sum_{m=2}^{L-1}\max_{r=\overline{0,P-1}}(K_{m,r}t^r(2n_{m-1}+3))+\\
&+(\sum_{m=2}^{L-1}(3+2n_{m+1})n_{m-1}n_m+5n_Ln_{L-1}+2n_L+3)\times t^0)),\\
Q_c = U\times(I\times(2L\times P^2-&2L\times P-3P^2+4P)+T\times(1+L\times P^2-L\times P-2P^2+2P)),
\end{aligned}\right\} \qquad (7)$$

с топологией передачи данных «звезда» как

$$\left.\begin{aligned}
Q_p = U\times(I\times(&\sum_{m=2}^{L}\max_{r=\overline{0,P-1}}(K_{m,r}t^r(2n_{m-1}+3))+\sum_{m=2}^{L-1}\max_{r=\overline{0,P-1}}(K_{m,r}t^r(3+2n_{m+1})n_{m-1})+\\
&+5\max_{r=\overline{0,P-1}}(K_{L,r}t^r n_{L-1}))+T\times(K_{2,r}t^r(2n_1+3)+\\
&+(\sum_{m=3}^{L}(2n_{m-1}+3)\times n_m+\sum_{m=2}^{L-1}(3+2n_{m+1})n_{m-1}n_m+5n_Ln_{L-1}+2n_L+3)\times t^0)),\\
Q_c = U\times(I\times(&2L\times P-L-P)+T\times(L\times P-2P-L+3)),
\end{aligned}\right\} \qquad (8)$$

и с топологией передачи данных «решетка» как

$$\left.\begin{aligned}
Q_p = U\times(H_j\times(&\max(H_{2,r}t^r(2n_1+3))+\max(H_{2,r}t^r(3+2n_3)n_1)+\\
&+\max(\max(H_{3,r}t^r(2n_2+3)),\max(H_{2,r}t^r(2n_1+3)))+\\
&+\sum_{m=3}^{L-1}\max(\max(H_{m,r}t^r(2n_{m-1}+3)),\max(H_{m+1,r}t^r(2n_m+3)))+\\
&+\max(\max(H_{L-1,r}t^r(2n_{L-2}+3)),\max(H_{L,r}t^r(2n_{L-1}+3))+5\max(H_{L,r}t^r n_{L-1}))+\\
&+\max(\max(H_{L,r}t^r(2n_{L-1}+3))+5\max(H_{L,r}t^r n_{L-1}),\max(H_{L-1,r}t^r(3+2n_L)n_{L-2}))+\\
&+\sum_{m=L-1}^{3}\max(\max(H_{L-2,r}t^r(3+2n_{L-1})n_{L-3}),\max(H_{L-1,r}t^r(3+2n_L)n_{L-2}))+\\
&+\max(\max(H_{2,r}t^r(3+2n_1)n_3),\max(H_{3,r}t^r(3+2n_2)n_4)))+\\
&+T(\sum_{m=2}^{L}\max(H_{m,r}t^r(2n_{m-1}+3))+\sum_{m=2}^{L-1}((3+2n_{m+1})n_{m-1}n_m)+5n_{L-1}n_L)),\\
Q_c = U\times(&I/b\times(\frac{13P-14}{2})+T\times(\frac{LP}{2}+1)),
\end{aligned}\right\} \qquad (9)$$

где z – количество ребер в топологии «решетка», которые идентифицируются номерами $j = \overline{1, z}$,

$K_{l,r}$ – критерий равномерного распределения по вычислителям нейронов каждого слоя определяется как

$$K_{l,r} = \begin{cases} \left\{(r-1)\left(\left[{n_l}/{P-1}\right]+1\right)+\varphi\right\}, \ \varphi = \overline{1, \left(\left[{n_l}/{P-1}\right]+1\right)}, \ 1 \le r \le r_d, \\ \left\{(P-b)\left(\left[{n_l}/{P-1}\right]+1\right)+(r-P+b-1)\left[{n_l}/{P-1}\right]+\varphi\right\}, \ \varphi = \overline{1, \left[{n_l}/{P-1}\right]}, \ r > r_d, \end{cases} \quad (10)$$

где $r_d = P - b$ – номер вычислителя, начиная с которого уменьшается на единицу значение критерия равномерного распределения параллельных операций K_r,

$b = n_l \times \mathrm{mod}(P-1)$ – количество вычислителей с большей нагрузкой,

$H_{l,r}$ – критерий равномерного распределения по вычислителям нейронов каждого слоя в топологии «решетка» определяется как

$$H_{l,r} = \begin{cases} \left\{r \times \left(\left[{zn_l}/{P}\right]+1\right)+\varphi\right\}, \ \varphi = \overline{1, \left(\left[{zn_l}/{P}\right]+1\right)}, \ 0 \le r < r_d, \\ \left\{(\frac{P}{z}-b)\left(\left[{zn_l}/{P}\right]+1\right)+\left(r-\frac{P}{z}+b\right)\left[{zn_l}/{P}\right]+\varphi\right\}, \ \varphi = \overline{1, \left[{zn_l}/{P}\right]}, \ r \ge r_d, \end{cases} \quad (11)$$

где $r_d = {P}/{z} - b$ – номер вычислителя, начиная с которого уменьшается на единицу значение критерия равномерного распределения параллельных операций K_r,

$b = n_l \times \mathrm{mod}\, {P}/{z}$ – количество вычислителей с большей нагрузкой,

H_j – текущие значения критерия в зависимости от номера ребра определяются как

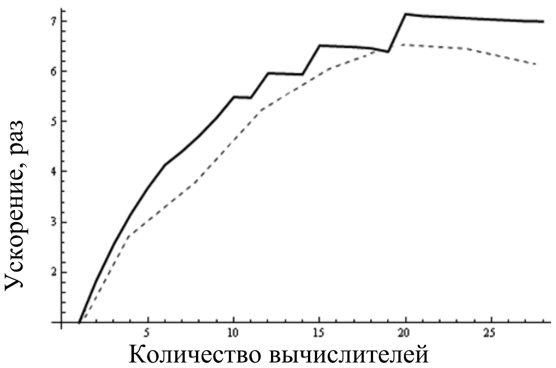
$$H_j = \begin{cases} \left\{(j-1)\left(\left[{I}/{z^2}\right]+1\right)+\varphi\right\}, \ \varphi = \overline{1, \left(\left[{I}/{z^2}\right]+1\right)}, \ 1 \le j \le r_d, \\ \left\{(z-b)\left(\left[{I}/{z^2}\right]+1\right)+(j-z+e-1)\left[{I}/{z^2}\right]+\varphi\right\}, \ \varphi = \overline{1, \left[{I}/{z^2}\right]}, \ j > r_d, \end{cases} \quad (12)$$

где $r_d = z - b$ – номер ребра, начиная с которого уменьшается на единицу значение критерия равномерного распределения обучающей выборки I,

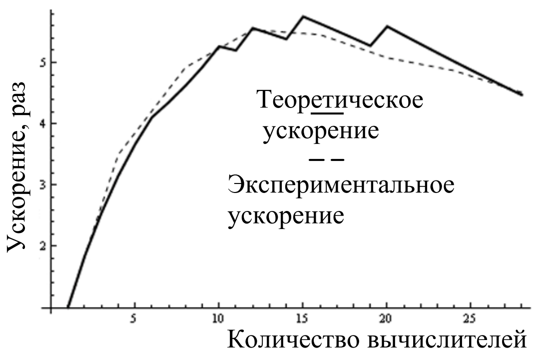
$b = I \times \mathrm{mod}\, z$ – количество ребер с большей нагрузкой.

На основе моделей производительности получены наиболее эффективные процедуры распределенного обучения многослойной нейронной сети методом обратного распространения ошибки с топологиями передачи данных «звезда», «полносвязный граф» и «решетка» [5]. На рис. 1 представлены результаты полученных на основании моделей (7), (8), (9) теоретических ускорений по сравнению с экспериментальными.

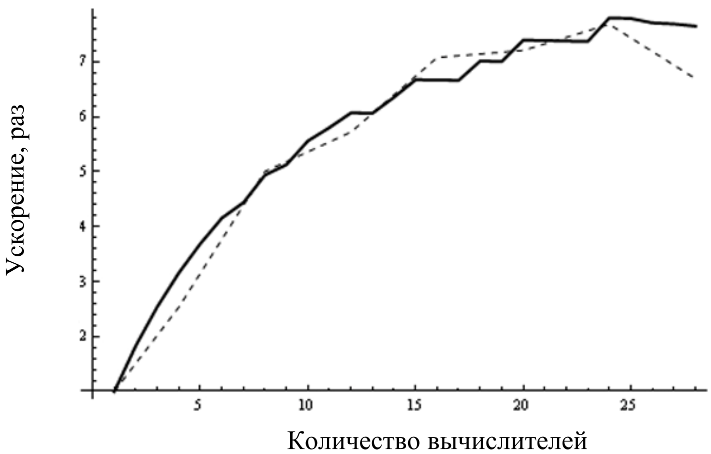
Для проведения экспериментальных исследований была решена задача классификации изображений размерностью 500×500 пикселей. В соответствии с выкладками в [6] были выбраны следующие исходные данные: L = 3 – количество слоев в многослойной нейронной сети (включая первый и выходной); n_1 = 250000, n_2 = 120, n_3 = 5 – количество нейронов в первом, втором и третьем слое соответственно; обучающая и тестовые выборки величиной I $= 30 \times 10^4$ и T = 3000 примеров; количество эпох U $= 90 \times 10^4$; количество вычислителей $P = \overline{1, 28}$.



а) Топология сети передачи данных «звезда»



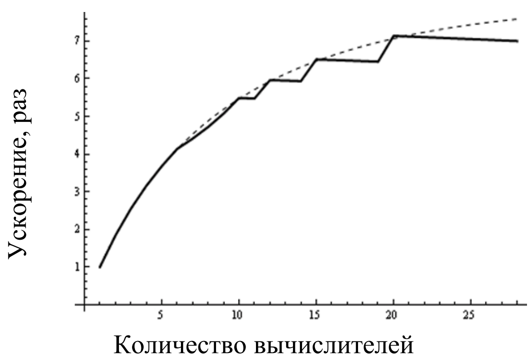
б) Топология сети передачи данных «полносвязный граф»



в) Топология сети передачи данных «решетка»

Рисунок 1 – Зависимость времени обучения нейронной сети от количества вычислителей

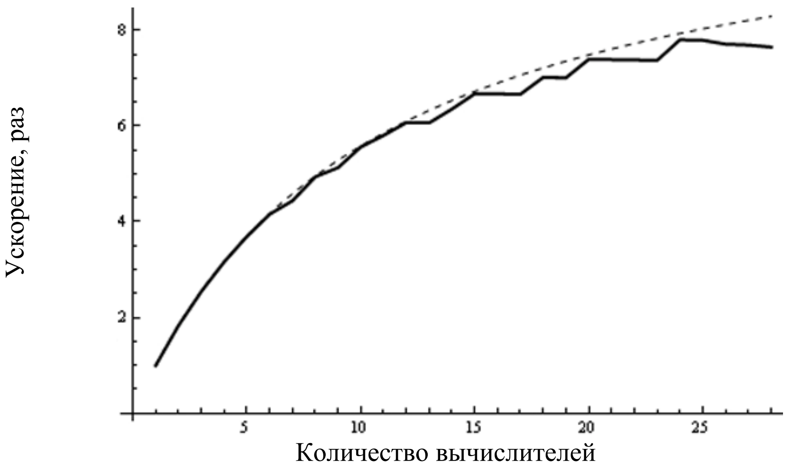
На рис. 2 показана графическая интерпретация методов определения ускорения с помощью полученного метода оценивания эффективности (3) и традиционного закона Амдала (1).





а) Топология сети передачи данных «звезда»

б) Топология сети передачи данных «полносвязный граф»



в) Топология сети передачи данных «решетка»

Рисунок 2 – Графическая интерпретация методов определения ускорения

Как видно из рис. 1, интерпретация теоретического ускорения, полученного на основании параметрических моделей (7), (8), (9), соответствует экспериментальным результатам, что говорит о достоверности использования предложенной модели для описания производительности распределенной процедуры обучения L-слойной нейронной сети. На рис. 2 отражено различие между графическими интерпретациями методов оценивания ускорения обучения L-слойной нейронной сети. На основании чего можно сделать вывод о том, что предложенный метод описывает ускорение распределенной процедуры в зависимости от значений задержки при барьерной синхронизации.

Выводы

Научная новизна работы заключается в том, что предложенные методы оценивания эффективности работы распределенных процедур при параллелизме на уровне задач и на уровне данных с помощью разработанных критериев равномерного распределения параллельных операций позволяют получить максимальное ускорение при наиболее эффективном использовании доступного количества вычислителей.

Разработанные параметрические модели производительности процедуры распределенного обучения многослойной нейронной сети в виде алгоритмических составляющих позволяют значительно сократить время ее обучения в распределенных компьютерных сетях с различными топологиями («решетка», «полносвязный граф», «звезда»).

Литература

1. Воеводин Вл.В. Решение больших задач в распределенных вычислительных средах / Вл.В. Воєводин //Автоматика и Телемеханика. – 2007. – № 5. – С. 32-45.
2. Amdahl G. Validity of the single-processor approach to achieving large-scale computing capabilities / G. Amdahl // Proc. 1967 AFIPS Conf., AFIPS Press. – 1967. – V. 30. – P. 483.
3. Гергель В.П. Основы параллельных вычислений для многопроцессорных вычислительных систем : [учебное пособие] / В.П. Гергель, Р.Г. Строгин. – Нижний Новгород : Изд-во ННГУ им. Н.И. Лобачевского, 2003. – 184 с.

4. Руденко О.Г. Искусственные нейронные сети: Учебное пособие / О.Г. Руденко, Е.В. Бодянский. – Харьков : ООО «Компания СМИТ», 2005. – 408 с.
5. Аксак Н.Г. Процедура параллельного обучения многослойной нейронной сети. Топология передачи данных «звезда» / Н.Г. Аксак, А.Ю. Лебёдкина, О.В. Хоменко // Науковий вісник Чернівецького національного університету імені Юрія Федьковича. Серія: Комп'ютерні системи та компоненти. – Том 1, випуск 2. – Черновцы:ЧНУ, 2010. – С. 95-103.
6. Ососков Г.А. Современные методы обработки экспериментальных данных в физике высоких энергий / Г.А. Осоков, А. Полянский., И.В. Пузынин // Научный обзорный журнал «Физика элементарных частиц и атомного ядра (ЭЧАЯ)». – Том 33, выпуск 3. – Дубна : ОИЯИ, 2002.– С. 676-745.

Literatura

1. VoevodinVl.V. Avtomatika iT elemehanika. № 5. 2007. S. 32-45.
2. Amdahl G. Proc. 1967 AFIPS Conf. AFIPS Press. V 30. 1967. P 483
3. Gergel' V.P. Osnovy parallel'nyh vychislenij dlja mnogoprocessornyh vychisl'nyh sistem. Nizhnij Novgorod : Izd-vo NNGU im. N.I. Lobachevskogo. 2003. 184 s.
4. Rudenko O.G. Iskusstvennye nejronnye seti: Uchebnoe posobie. Har'kov: OOO "Kompanija SMIT". 2005. 408 s.
5. Aksak N.G. Naukovyj visnyk Chernivec'koho nacional'noh ouniversytetu imeni Yuriya Fed'kovycha. Seriya: Komp'yuterni systemy ta komponenty. Tom 1. Vypusk 2.Chernovcy :ChNU. 2010. S 95-103
6. Ososkov G.A. Nauchnyj obzornyj zhurnal "Fizika jelementarnyh chastic i atomnogo jadra (JeChAJa)". Tom 33. Vypusk 3. Dubna: OIJaI. 2002. S 676-745

Н.Г. Аксак, А.Ю. Лебьодкіна

Методи і моделі продуктивності навчання багатошарових нейронних мереж в розподілених комп'ютерних середовищах

У статті запропоновано методи та моделі продуктивності паралельних процедур, які дозволяють ефективно застосовувати багатошарову нейронну мережу в розподілених комп'ютерних середовищах з різними топологіями («решітка», «повнозв'язний граф», «зірка»). Достовірність використання запропонованих методів і моделей підтверджується експериментальними дослідженнями.

N.G. Axak, A.U. Lebodkina

Methods and Performance Models of Training Multilayer Neural Networks in Distributed Computing Environments

The methods and performance modelsof parallel processes that enable effectivemultilevel neural networks use in distributed computing environments with different topologies ("grid", "fully connected graph", "star") are proposed inthe paper. The reliability of the proposed methods and models is confirmed by experimental researches.

Статья поступила в редакцию 08.07.2011.